PENERAPAN ALGORITME C4.5 UNTUK KLASIFIKASI KASUS COVID-19

Sirojul Alam
Universitas Buana Perjuangan
Karawang, Indonesia
sirojmu@gmail.com

Amril Mutoi Siregar
Universitas Buana Perjuangan
Karawang, Indonesia
amrilsiregar@ubpkarawang.ac,id

Ayu Ratna Juwita
Universitas Buana Perjuangan
Karawang, Indonesia
ayu.rj@ubpkarawang.ac.id

***Abstract*— Wabah virus corona versi SARS-nCov2 telah terjadi di akhir tahun 2019 di Wuhan, Tiongkok. Karena kemampuan penularan yang supercepat, virus ini sudah bertransmisi lintas negara. Hampir semua negara terkena wabah virus ini, hingga PBB melalui WHO menjadikan wabah virus ini sebagai pandemi global. Penelitian ini menggunakan teknik klasifikasi data mining dengan algoritme C4.5 dengan alat bantu Orange data mining. Juga dilakukan pengolahan dengan bahasa pemrograman Python menggunakan library standar yang sudah disediakan seperti matplotlib, seaborn, numpy, pandas, dan decisionTree Classification. Klasifikasi dilakukan untuk mengelompokkan data kasus terkonfirmasi, meninggal, dan sembuh dari covid 19 menjadi tinggi dan rendah. Hasil klasifikasi didapatkan nilai AUC 0.871, yang menunjukkan klasifikasi kategori baik, nilai accuracy sebesar 93%, nilai precision sebesar 95%, dan nilai recall 96%. Pada pemodelan menggunakan Python, nilai akurasi yang didapatkan adalah 94%, nilai presisi adalah 98%, dan nilai recall yang didapatkan adalah 95%. Penelitian juga berhasil memprediksi tren kasus di Indonesia dengan menggunakan model fungsi logistic, model standar yang digunakan untuk memprediksi laju pertumbuhan populasi. Hasil prediksi yang didapatkan adalah rata-rata laju pertumbuhan virus adalah 0.0255, puncak pandemi terjadi pada hari ke-196 sampai 200 atau tanggal 14 sampai 18 September 2020, akhir pandemi diprediksi pada hari ke-717 atau tanggal 17 Februari 2022 mendatang.**

Kata kunci — covid19, data mining, klasifikasi, pandemi, , prediksi, python

I. PENDAHULUAN

Covid-19 merupakan varian baru virus Corona yang awalnya belum pernah diidentifikasi dapat menyerang manusia [1]. Kasus awal virus ini menyerang manusia terjadi di Provinsi Wuhan, Tiongkok. Semula hanya didiagnosa sebagai pnumonia biasa dengan gejala yang mirip dengan sakit flu yang umum seperti batuk, demam, sesak nafas dan nafsu makan menurun; akan tetapi ternyata virus ini berkembang dengan cepat sehingga dapat menyebabkan gagal organ. Kondisi akan semakin parah jika pasien mempunyai penyakit bawaan. Karena penularan yang supercepat ini, induk organisasi kesehatan dunia atau WHO menjadikan wabah Covid-19 ini sebagai pandemi dipenghujung tahun 2019. Penetapan pandemi menjadi tanda bahwa wabah Covid-19 telah menyerang hampir semua negara di dunia serta memberikan tantangan tersendiri untuk meramu cara yang efektif menyelesaikannya serta cara yang paling jitu untuk meminimalisir efek yang ditimbulkannya terutama bidang ekonomi dan kesehatan karena dua bidang tersebut termasuk yang utama dalam keberlangsungan hidup manusia.

Para peneliti di seluruh dunia sedang berlomba-lomba meneliti tentang Covid-19. Semua sadar akan efek domino yang ditimbulkan dari pandemi ini. Para ahli di bidang pendidikan meneliti efektifitas sistem pembelajaran dimasa pandemi, para ahli di bidang epidemi meneliti kapan puncak pandemi terjadi, para ahli di bidang bisnis dan manajemen meneliti bagaimana cara mempertahankan bisnis di tengah pandemi, para ahli virologi berlomba menemukan vaksin Covid-19 yang aman untuk manusia termasuk para ahli di Indonesia. Para ahli dibidang teknologi informasi juga tidak ketinggalan, dengan menggunakan kecanggihan teknologi komputasi saat ini, teknologi informasi mampu menghadirkan solusi tak langsung dalam penanganan pandemi Covid-19. Berbagai cara dilakukan untuk menghasilkan metode penanganan terbaik berdasarkan data kasus yang terjadi.

Data mining yang merupakan salah satu cabang teknologi informasi dapat digunakan untuk memprediksi; mengklasifikasi; maupun mengelompokkan kasus Covid-19. Penggunaan teknik-teknik ini diharapkan mampu menghasilkan model yang efektif guna menangani dan mengendalikan pandemi yang semakin meluas. Beberapa teknik data mining terkenal adalah klasifikasi, klastering, asosiasi, peramalan, dan estimasi [2]. Teknik klasifikasi adalah salah satu teknik yang sering digunakan [3], [4], [5]. Sedangkan pada teknik klasifikasi, algoritma C4.5 merupakan yang paling sering digunakan [6], [7], [8].

II. TINJAUAN PUSTAKA

A. *Klasifikasi*

Klasifikasi untuk kali pertama dikenalkan oleh Corolus von Linne, sebagai orang pertama yang melakukan klasifikasi terhadap tanaman dengan spesies tertentu [9]. Klasifikasi merupakan metode yang digunakan untuk membedakan kelas data yang labelnya belum diketahui [10]. Aturan yang terbentuk berupa model "if-then", pohon keputusan atau decision tree, maupun neural network. Klasifikasi bertujuan untuk mendapatkan model pohon keputusan yang dapat digunakan sebagai media prediksi serta melihat keterkaitan antar variabel dalam suatu data [11]. Menurut Defiyanti dan Jajuli, klasifikasi merupakan cara menemukan model atau fungsi sebagai representasi kelas data untuk kepentingan tertentu [12], dengan klasifikasi kita dapat menentukan kelas label dari suatu data [13].

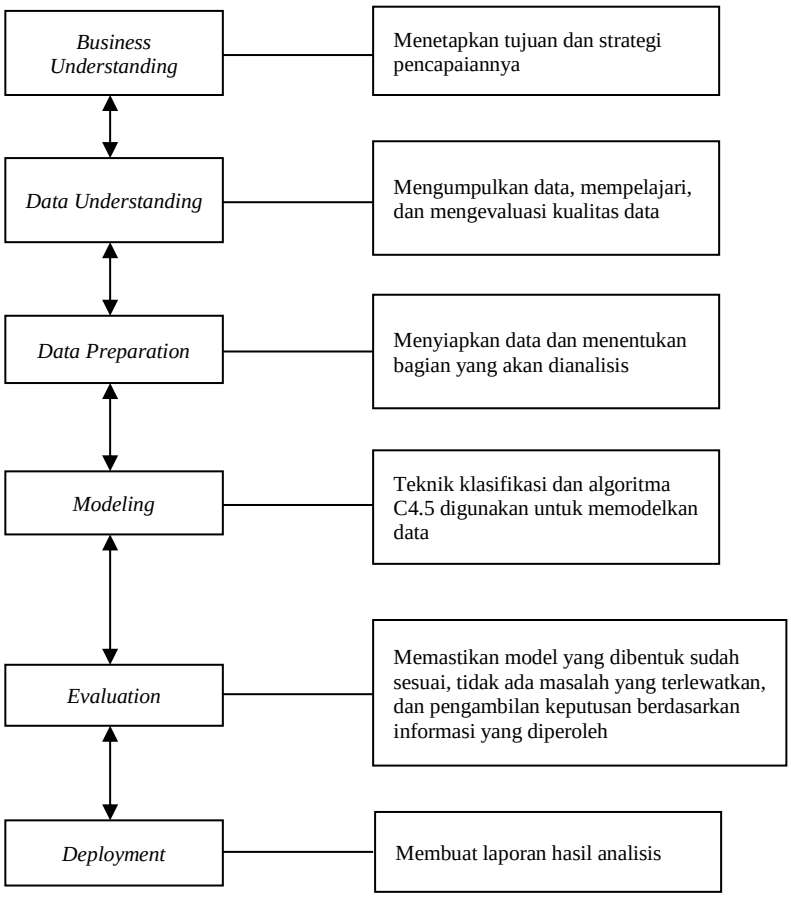
B. *Algoritme C4.5*

Algoritma C4.5 merupakan pengembangan dan penyempurnaan algoritma decision tree ID3 atau Iterative Dichotomizer yang dibuat oleh J. Ross Quinlan diakhir 1970 sampai awal 1980 [14]. Kelebihan algoritma ini adalah pada fleksibilitas dan mudah dimengerti serta dapat divisualisasi dalam gambar pohon keputusan [15]. Pohon keputusan yang dibentuk oleh algoritma C4.5 terdiri akar untuk atribut yang paling atas dan daun untuk atribut paling bawah. Algoritma C4.5 mengklasifikasikan data ke dalam kelasnya masing-masing [16]. Algoritma ini sangat kuat dan terkenal dalam teknik klasifikasi dan prediksi [17].

C. *Covid-19*

Covid-19 adalah akronim dari Corona Virus Disease yang ditemukan pada tahun 2019. Saat pertama kali diidentifikasi virus ini dinamakan SARS-nCov atau varian baru virus corona yang juga menyebabkan penyakit gangguan pernafasan SARS yang pernah mewabah beberapa tahun lalu. Induk organisasi kesehatan dunia atau World Health Organization (WHO) sepakat untuk menyebut virus ini dengan nama Covid-19. Saat ini sekitar 196 negara telah terjangkiti Covid-19. Pada tanggal 30 November 2020 jumlah total kasus Covid-19 diseluruh dunia sudah diangka 62.7 juta kasus, jumlah yang sembuh 40.1 juta dan jumlah yang meninggal dunia mencapai 1.46 juta. Di Indonesia sendiri pada tanggal 30 November 2020 jumlah kasus terkonfirmasi positif Covid-19 mencapai 534ribu, sedangkan jumlah yang sembuh mencapai 446ribu, dan jumlah yang meninggal mencapai 16.815, dengan jumlah kasus terbanyak ada di ibu kota Jakarta (www.google.com). Jumlah kasus tersebut akan terus naik karena belum ditemukannya vaksin untuk virus ini. Beberapa cara dilakukan untuk mencegah meluasnya wabah ini diantaranya kampanye untuk selalu mengenakan masker jika bepergian keluar rumah, mencuci tangan dengan sabun di air yang mengalir, serta selalu manjaga jarak dan menghindari kerumunan. Dengan melakukan 3 hal tersebut diharapkan masyarakat tidak menjadi penular maupun tertular virus [18].

III. METODE PENELITIAN



Berdasarkan Gambar 1, penelitian dimulai dengan penetapan tujuan penelitian dan strategi untuk mencapai tujuan penelitian tersebut, keduanya masuk dalam fase pertama penelitian yaitu Business Understanding. Kemudian penelitian dilanjutkan dengan mengumpulkan data melalui online repository https://www.kaggle.com. Di kaggle.com ini, peneliti mendapatkan data kasus pandemi Covid-19 di sebagian besar negara di dunia. Setelah didapatkan, data akan dipelajari lebih lanjut untuk mengenali data dan dilakukan evaluasi dengan menghilangkan missing value untuk mendapatkan data yang baik untuk proses mining. Proses-proses ini masuk dalam fase kedua penelitian yaitu Data Understanding. Setelah didapatkan data yang baik untuk proses mining, langkah selanjutnya adalah menyiapkannya dan menentukan bagian data mana yang akan

dilakukan analisis. Menentukan variabel yang menjadi acuan, yaitu confirmed, deaths, dan recovered, kemudian nilai ketiga variabel tersebut dikategorikan menjadi tinggi, untuk nilai yang berada diatas rata-rata kasus, dan rendah untuk nilai yang berada dibawah rata-rata kasus. Fase ini adalah fase ketiga penelitian yant disebut Data Preparation. Fase penelitian keempat adalah fase Modeling, dimana pada fase ini peneliti menentukan teknik dan algoritma apa yang akan digunakan. Peneliti memutuskan untuk menggunakan teknik klasifikasi dan algoritma C4.5 pada proses data mining ini. kemudian dievaluasi untuk mengetahui kategori klasifikasi dan kategori modelnya. Deployment adalah fase terakhir dari penelitian. Fase ini berupa pembuatan laporan hasil penelitian berupa karya tulis atau jurnal ilmiah. Keenam fase penelitian tersebut dilakukana secara berkesinambungan, antara fase yang satu saling berkaitan dengan fase yang lain, sehingga jika suatu fase penelitian belum selesai maka fase yang lain tidak dapat dilakukan.

IV. HASIL DAN PEMBAHASAN

Penelitian dilakukan dengan memodelkan kasus sesuai dengan algoritma yang akan digunakan, yaitu algoritma klasifikasi C4.5. Algoritma ini digunakan untuk mengklasifikasikan kasus Covid-19 kedalam kelas terkonfirmasi. Gambar 2 adalah proses yang dilakukan menggunakan bantuan aplikasi Orange Data mining.

Gambar . Proses data mining dengan aplikasi Orange

File dataset covid-19 berformat .csv di import kedalam area kerja kemudian dataset diperiksa dengan widget Data Table. Widget adalah seperangkat alat dengan fungsi tertentu pada Orange Data Mining. Selanjutnya setelah dataset dipastikan tidak ada kesalahan, dan sudah sesuai dengan tujuan pengolahan, algoritma C4.5 digunakan dalam pemodelan. Setelah selesai algoritma C4.5 diuji dengan menggunakan widget Test and Score. Dengan widget ini peneliti mendapatkan nilai AUC, accuracy, precision dan recall atau sensitivity. Nilai AUC adalah nilai yang didapatkan dari kurva ROC (Recheiver Operating Characteristics). Selain merupakan ekpresi dari confusion matrix [13], ROC adalah cara yang dapat digunakan untuk mengukur tingkat klasifikasi [19], dan dari ROC ini akan didapatkan nilai AUC atau Area Under ROC Curve sebagai acuan kategori klasifikasi. Berikut adalah rentang nilai AUC yang dimaksud:

0.90-1.00 = Klasifikasi sangat baik

0.80-0.90 = Klasifikasi baik

0.70-0.80 = Klasifikasi cukup

0.60-0.70 = Klasifikasi buruk

0.50-0.60 = Klasifikasi salah

Kemudian untuk mengetahui tabel probabilitas yang terbentuk, peneliti menggunakan widget Confusion Matrix. Gambar 3 berikut adalah *confusion matrix.*

		Predicted class	
		Yes	No
Actual class	Yes	TP	FN
	No	FP	TN

Berdasarkan Gambar 3 diatas, terdapat empat kemungkinan yang kondisi yang terjadi; jika actual class dan predicted class bernilai positif maka kondisinya disebut true positive (TP), kondisi dimana data yang bernilai positif diprediksi benar oleh sistem. Jika actual class bernilai positif kemudian diprediksi salah oleh sistem maka kondisinya disebut false negative (FN). Kemudian jika actual class bernilai negatif dan diprediksi positif oleh sistem maka kondisinya disebut sebagai false positive (FP), dan jika actual class bernilai negatif kemudian dinilai negatif juga oleh sistem maka kondisi tersebut dinamakan true negative (TN). Nilai *accuracy, precision,* dan *recall* didapatkan dari tabel probabilitas atau confusion matrix [20], [19]. Nilai accuracy mendeskripsikan kemampuan sistem untuk secara akurat mengklasifikasi data dengan benar. Sedangkan nilai *precision* adalah hasil perbandingan antaran jumlah data positif yang diklasifikasi secara benar dengan total data klasifikasi

positif, dan nilai *recall* menunjukkan persentase data berkategori positif yang diklasifikasi dengan benar oleh sistem. Berikut adalah formula perhitungan nilai *accuracy*, *precision*, dan *recall* berdasarkan confusion matrix.

$$Accuracy = (TP + TN)/(TP + TN + FN + FP) \quad (1)$$

$$Precision = TP/(TP + FP) \quad (2)$$

$$Recall = TP/(TP+FN) \quad (3)$$

Formula-formula tersebut selanjutnya akan digunakan pada fase berikutnya, yaitu fase evaluation, fase dimana peneliti mengevaluasi hasil pemodelan pada fase sebelumnya. Peneliti juga melakukan pemodelan menggunakan bahasa pemrograman Python untuk mengukur tingkat *accuracy*, *precision*, dan *recall* menggunakan *library DecisionTreeClassifier*. Sebelum dilakukan pemodelan, dataset diatur sedemikian rupa sehingga dapat diolah dengan baik. Pengaturan dataset yang dilakukan diantaranya adalah merubah kolom Waktu yang berisi Pagi, Siang, Sore, dan Malam menjadi kolom dengan nama Kategori dengan nilai 1 untuk pagi, 2 untuk siang, 3 untuk sore, dan 4 untuk malam. Pengaturan lainnya adalah merepresentasikan kasus menjadi 0 jika rendah dan 1 jika tinggi, sehingga nilai baru pada kolom *Deaths*, *Recovered*, dan *Confirmed* menjadi 0 atau 1.

	ObservationDate	Province_State	Country_Region	Waktu	Kategori	Deaths	Recovered	Confirmed
0	January	Anhui	Mainland China	sore	3	0	0	0
1	January	Beijing	Mainland China	sore	3	0	0	0
2	January	Chongqing	Mainland China	sore	3	0	0	0
3	January	Fujian	Mainland China	sore	3	0	0	0
4	January	Gansu	Mainland China	sore	3	0	0	0

Gambar 5. Dataset

Kemudian dilakukan pemisahan *feature* data dengan label kelas data dan menentukan data latih dan data uji. Para pemodelan ini peneliti menggunakan data uji sebanyak 20% atau 0.20 dari total data dan membuat model klasifikasi seperti pada Gambar 5 dibawah ini.

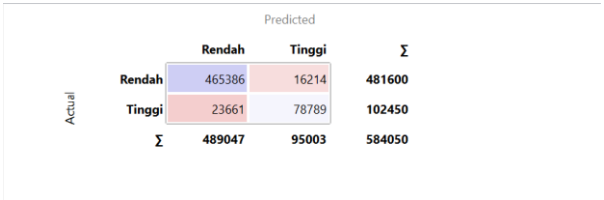
```
classifier = DecisionTreeClassifier()
classifier.fit(X_train, y_train)

DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='gini',
                       max_depth=None, max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, presort='deprecated',
                       random_state=None, splitter='best')
```

Gambar 6. Klasifikasi DecisionTree

Setelah mendapatkan model klasifikasi seperti pada Gambar 5, proses selanjutnya adalah menyimpan hasil prediksi dan mengukur akurasi menggunakan *confusion matrix*.

Hasil penelitian menggunakan bantuan aplikasi Orange data mining dan Python ditampilkan dalam confusion matrix dibawah ini.



Gambar 7. Confusion Matrix dengan Orang data mining

Berdasarkan Gambar 6 di atas, kategori klasifikasi yang peneliti lakukan termasuk dalam kategori yang baik dengan nilai AUC 0.872. Nilai *accuracy* yang didapatkan adalah 0.932, nilai *precicion* yang didapatkan adalah 0.930, dan nilai *recall* yang didapatkan adalah 0.932, serta nilai F1, nilai yang mendeskripsikan perbandingan *precision* dan *recall*, adalah 0.931. Nilai F1 digunakan jika dataset yang digunakan memiliki jumlah FP dan FN yang tidak simestris, dan jika jumlah FP dan FN dalam dataset yang digunakan adalah simetris maka *accuracy* cukup sebagai acuan untuk mengukur performa algoritma [21]. Nilai-nilai tersebut cukup menggambarkan bahwa klasifikasi yang dilakukan sudah baik dengan nilai *accuracy*, *precision*, dan *recall* yang terbilang bagus karena lebih dari 90%. Nilai-nilai tersebut juga didapatkan pada confusion matrix dengan menggunakan Python seperti yang terlihat pada Gambar 7 berikut.

```
[[18179  1055]
 [  312  3815]]
              precision    recall  f1-score   support

           0       0.98      0.95      0.96     19234
           1       0.78      0.92      0.85      4127

    accuracy                           0.94     23361
   macro avg       0.88      0.93      0.91     23361
weighted avg       0.95      0.94      0.94     23361
```

Gambar 8. Confusion Matrix dengan Python

Dari Gambar 7 dapat diketahui nilai *Accuracy, Precision*, dan *Recall* yang peroleh dengan menggunakan bahasa pemrograman Python. Dengan Python penulis mendapatkan nilai Accuracy sebesar 0.94 atau 94%, dan nilai precision sebesar 0.98 atau 98%, sedangkan untuk nilai recall adalah 0.95 atau 95%.

V. KESIMPULAN

Dengan bantuan Orang data mining, nilai accuracy sebesar 93%, nilai precision sebesar 95%, dan nilai recall sebesar 96%, dengan nilai AUC sebesar 0.872. Dengan pemodelan menggunakan bahasa pemrograman Python dengan memanfaatkan beberapa library pengolahan data yang sudah tersedia, nilai accuracy yang didapatkan adalah 0.94, sedangkan nilai precision-nya adalah 0.98, dan nilai recall yang didapatkan adalah 0.95. Penelitian selanjutnya dapat dilakukan dengan menggunakan algoritma klasifikasi yang lain seperti naïve bayes dan CART.

PENGAKUAN

Makalah ini adalah sebagian dari penelitian Tugas Akhir milik Sirojul Alam dan disponsori oleh Penerapan Algoritme C4.5 untuk Klasifikasi Kasus Covid-19.

DAFTAR PUSTAKA

[1] World Health Organization, “Laboratory Guidelines for the Detection and Diagnosis of COVID-19 Virus Infection,” *Paho*. 2020.

[2] I. Kamila, U. Khairunnisa, and M. Mustakim, “Perbandingan Algoritma K-Means dan K-Medoids untuk Pengelompokan Data Transaksi Bongkar Muat di Provinsi Riau,” *J. Ilm. Rekayasa dan Manaj. Sist. Inf.*, 2019, doi: 10.24014/rmsi.v5i1.7381.

[3] J. Jamal and D. Yanto, “Analisis RFM dan Algoritma K-Means untuk Clustering Loyalitas Customer,” *Energy*, 2019.

[4] D. Marlina, N. Lina, A. Fernando, and A. Ramadhan, “Implementasi Algoritma K-Medoids dan K-Means untuk Pengelompokkan Wilayah Sebaran Cacat pada Anak,” *J. CoreIT J. Has. Penelit. Ilmu Komput. dan Teknol. Inf.*, 2018, doi: 10.24014/coreit.v4i2.4498.

[5] E. H. S. Atmaja, “Implementation of k-Medoids Clustering Algorithm to Cluster Crime Patterns in Yogyakarta,” *Int. J. Appl. Sci. Smart Technol.*, 2019, doi: 10.24071/ijasst.v1i1.1859.

[6] W. Katrina, H. J. Damanik, F. Parhusip, D. Hartama, A. P. Windarto, and A. Wanto, “C.45 Classification Rules Model for Determining Students Level of Understanding of the Subject,” 2019, doi: 10.1088/1742-6596/1255/1/012005.

[7] D. Hartama, A. Perdana Windarto, and A. Wanto, “The Application of Data Mining in Determining Patterns of Interest of High School Graduates,” 2019, doi: 10.1088/1742-6596/1339/1/012042.

[8] M. Widyastuti, A. G. Fepdiani Simanjuntak, D. Hartama, A. P. Windarto, and A. Wanto, “Classification Model C.45 on Determining the Quality of Custumer Service in Bank BTN Pematangsiantar Branch,” 2019, doi: 10.1088/1742-6596/1255/1/012002.

[9] Y. Mardi, “Data Mining : Klasifikasi Menggunakan Algoritma C4.5,” *J. Edik Inform.*, vol. 2, no. 2, pp. 213–219, 2017.

[10] P. Assiroj, “Kajian Perbandingan Teknik Klasifikasi AlgoritmaC4.5, Naive Bayes, dan CART untk Prediksi Kelulusan Mahasiswa (Studi Kasus: STMIK ROSMA Karawang),” *Media Inform.*, vol. 15, no. 2, pp. 1–17, 2016, doi: 10.5281/zenodo.1184054.

[11] E. E. Pramadhani and T. Setiadi, “210945-Penerapan-Data-Mining-Untuk-Klasifikasi.Pdf.” 2014.

[12] S. Defiyanti and M. Jajuli, “Integrasi Metode Klasifikasi Dan Clustering dalam Data Mining,” no. March, pp. 39–44, 2015.

[13] S. Hendrian, “Algoritma Klasifikasi Data Mining Untuk Memprediksi Siswa Dalam Memperoleh Bantuan Dana Pendidikan,” *Fakt. Exacta*, vol. 11, no. 3, pp. 266–274, 2018, doi: 10.30998/faktorexacta.v11i3.2777.

[14] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. 2016.

[15] W. T. Ina, “Klasifikasi Data Rekam Medis Berdasarkan Kode Penyakit Internasional Menggunakan Algoritma C4.5,” *J. Media Elektro*, 2013.

[16] K. Hastuti, “Analisis komparasi algoritma klasifikasi data mining untuk prediksi mahasiswa non aktif,” *Semin. Nas. Teknol. Inf. Komun. Terap.*, 2012.

[17] H. D. Honesqi, "Klasifikasi Data Mining Untuk Menentukan Tingkat Persetujuan Kartu Kredit," *J. Teknoif*, vol. 5, no. 2, pp. 57–62, 2017, doi: 10.21063/jtif.2017.v5.2.57-62.
[18] A. F. Watratan, A. P. B, and D. Moeis, "Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid-19 Di Indonesia," *J. Appl. Comput. Sci. Technol. (Jacost)*, vol. 1, no. 1, pp. 7–14, 2020.
[19] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Elsevier B.V., 2012.
[20] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009, doi: 10.1016/j.ipm.2009.03.002.
[21] J. Xu, Y. Zhang, and D. Miao, "Three-way confusion matrix for classification: A measure driven view," *Inf. Sci. (Ny).*, vol. 507, no. July, pp. 772–794, 2020, doi: 10.1016/j.ins.2019.06.064.